DOCUMENT RESUME

ED 179 919 CS 005 143

AUTHOR Moe, Alden J.
TITLE Analyzing Text with Computers: The State of the Technology.
PUB DATE 79
NOTE 13p.; Paper presented at the Annual Meeting of the National Reading Conference (29th, San Antonio, TX, November 29-December 1, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Computer Programs; *Computers; *Discourse Analysis; Evaluation Methods; Literature Reviews; *Readability; Reading Materials; Reading Research; Structural Analysis; Syntax; *Vocabulary; Word Frequency

ABSTRACT

A review of the literature and an examination of the state of the technology is presented in this discussion of text analysis with computers. The computer capabilities of several available text analysis programs are reviewed, including those that examine vocabulary, word strings, sentence and syntactic complexity, and readability. (AEA)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED179919

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Alden J. Moe
205 Education Building
Purdue University
West Lafayette, IN 47907

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Alden J. Moe

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ANALYZING TEXT WITH COMPUTERS: THE STATE OF THE TECHNOLOGY[1]

Alden J. Moe

Purdue University

Computer technology is now at the stage where large amounts of text may be analyzed with both precision and speed. Among the types of text analysis for which computer programs are currently available are those which examine 1) vocabulary, 2) word strings, 3) sentence and syntactic complexity, and 4) readability.

The purpose of this paper is to provide a review of the literature and to examine the state of the technology insofar as text analysis with computers is concerned. The computer capabilities of several available text analysis programs will also be presented. The paper is organized according to the four major topics identified above.

---

[1]A paper presented at the annual meeting of the National Reading Conference, San Antonio, Texas, November 30, 1979.

CS005143

Vocabulary Analysis

Computers have been used for the counting and sorting of words for at least fifteen years (Kucera & Francis, 1967) and, assuming a sufficiently large memory, hundreds of thousands of running words of text may be ingested by card or tape and listed alphabetically and/or ranked by each word type according to frequency of use. Studies reported by Carroll, Davies, and Richman (1971) and Harris and Jacobson (1972) used corpuses of approximately five million and four-and-a-half million words respectively. Other reports (Durr, 1973; Moe, 1973; Hopkins, 1979; and Moe, Rush & Storlie, 1979) of vocabulary analysis on a smaller scale indicate that similar programs are available.

Programs developed for word counting and sorting often provide a listing of all different words and a percentage of use figure for each word. Some programs also provide type-token ratios, a measure of vocabulary diversity, and such use has also been reported (Fox, 1972; Moe, 1974; Hopkins, 1976; McCormick, 1976; Ratkowsky & Hantrais, 1975; and Woods & Moe, 1977).

A summary of the information provided by the vocabulary analysis program developed by Moe is listed below:

1. the total number of running words (tokens) in the text.
2. the total number of different words (types) in the text.
3. an alphabetical listing of all words used.
4. a sorted listing of words based on frequency of use.
5. a percentage of use figure for each word.
6. the average word length and standard deviation.
7. the number of words of one letter, two letters, etc.

8. the total number of sentences.
9. the average sentence length and standard deviation.
10. the number of sentences of one word, two words, etc.
11. the number of syllables per text passage.
12. a corrected type-token ratio per text passage.

The expertise necessary to write the programs which have the capabilities described above should be available at most computer centers. The only major limitation--at any computer facility--will be the size of the computer memory and, therefore, the amount of text analyzed in a single batch may be limited.

Word String Analysis

Word string analysis involves the identification and some kind of subsequent analysis (or comparison, perhaps) of word strings of two or more words which occur contiguously in text. String analysis has been conducted as one means of examining syntactic complexity (Selden, 1978) and for identifying the common strings--or in some cases phrases--in text (Moe & Hopkins, 1978).

As with vocabulary analysis, the major limitation occurs with computer memory size; even the most common two-word strings (in the, of the, to the, and the, and on the) occur relatively infrequently as compared to the individual common words (the, of, and, to, in, and on) which make up the string. Even with a corpus of moderate size, 100,000 words, for example, the memory limits which are routinely available may be exceeded.

Sentence and Syntactic Analysis

Sentence analysis and syntactic analysis are not necessarily the same unless one assumes that sentence length is an accurate measure of syntactic complexity,[1] e.g., an eight-word sentence is more syntactically complex than a seven-word sentence, etc. If that assumption is accepted, syntactic analysis is easily determined and the programming is straightforward; a marker such as a period signals the termination of a sentence (be they exclamatory, interrogative, or declarative).

Where syntactic complexity is determined by something other than sentence length such as the occurrence of minimal terminable units (T-units) or prepositional phrases, for example, then the programming becomes rather complex. It appears, however, that Golub and Kidder (1974; also see Kidder & Golub, 1976) have provided an accurate, elaborate measure of examining syntax, through T-units, which is wholly computer determined and their program has been used by at least one other language researcher (Hopkins, 1976). Pepinsky (1978) and Rush and others (1974) also appear to have developed a program for the syntactic analysis of text although Pepinsky recommends that after certain stages of the computer analysis "one or more humans edit the text..correcting the relatively few evident 'errors' made by computer" (p. 2).

[1] The use of readability formulas is based on such a premise and it may be that some of the newer measures of syntactic complexity are essentially measuring what is measured in sentence length. For an empirical study of such a comparison, see J. W. Miller & C. A. Hintzman, 1975.

A summary of the information provided by the Golub and Kidder program is found below:

1. the total number of words.
2. the total number of T-units.
3. the number of words per T-unit.
4. the number of subordinate clauses per T-unit.
5. the main clause word length (mean).
6. the subordinate clause word length (mean).
7. the number of modals.
8. the number of be and have forms in the auxiliary.
9. the number of prepositional phrases.
10. the number of possessive nouns and pronouns.
11. the number of adverbs of time.
12. the number of gerunds, participles, and absolute phrases.
13. a syntactic density score.
14. a grade level conversion table for the syntactic density score.

The development of computer programs in this area requires the close cooperation of the language specialist and the computer programmer; it helps if such expertise is found in one person.

Readability Estimation

The use of computers in estimating readability of text is no longer uncommon (Mason & Blanchard, 1979) and there are a number of locations[1] in the United State where such use is found. Watkins (1975),

---

[1] Among those locations are the following: Bell Laboratories (Murray Hill, N.J.), California State University at Hayward, Naval Education and Training Command (Pensacola), Oregon State University, Pensacola Junior College, Purdue University, San Diego State University, University of Minnesota, and the University of Virginia.

Berry (1979), and Rush (1979) report the use of programs which simultaneously calculate three or more readability estimates.

Most readability formulas use sentence length as one factor in the calculations. This factor is easily determined by computer and the average sentence length of a passage is easily obtained.

Readability formulas also use some index of vocabulary diversity which is based on a syllable count or on the number of words not found on someone's list of easy or familiar words (such as the Dale list, for example). Both of these measures, a syllable count or an unfamiliar word count, may be programmed for computer analysis.

Fang (1968), Coke and Rothkopf (1970), Felsenthal (1973), Coleman and Liau (1975), and Shamo (1975) all report the use of an accurate substitute for an actual syllable count. The substitutes usually involve a score or count based on vowels per word, consonants per word, letters per word, or a letters-per-syllable constant. In a validation study, Hopkins and Moe (1975) recommended the use of the Felsenthal constant of 3.1127 characters as being the equivalent of one syllable.

For readability formulas requiring the use of a word list, the list may be included in the program as an internal dictionary (Berry, 1979; Rush, 1979). Any word from the text sample not found in the computer's dictionary is then considered a hard word for the readability calculations. For formulas which count all forms of a word on the list of familiar words as familiar, a special routine may have to be included in the program to handle plurals, possessives of nouns, derivatives of verbs, etc.; if the computer system is large enough, however, the plurals, possessives of nouns, and the like may be in-

cluded as separate words in the computer's dictionary.

Many publishing companies now use computer type setting. With computerized readability formulas as part of the computer type setting system, it is possible to calculate estimates--using the entire text of the publication--for every piece of printed material produced in such a system.

Summary

The analysis of text by computers has been demonstrated to be both feasible and practical. Computers are readily available and the programs are rapidly improving. The versatility and the efficiency of computers for text analysis should convince educators that computers are enormously useful tools for studying language.

References

Barry, J. B. Computerized readability levels--their need and use. Journal of Educational Data Processing, 1979, 16, 10-22.

Barry, J. B., & Stevenson, T. E. Using a computer to calculate the Dale-Chall formula. Journal of Reading, 1975, 19(3), 218-222.

Carroll, J. B., Davies, P., & Richman, B. Word Frequency Book, Boston: Houghton Mifflin, 1971.

Coke, E. U., & Rothkopf, E. Z. Note on a simple algorithm for a computer produced reading ease score. Journal of Applied Psychology, 1970, 54, 209-210.

Coleman, M., & Liau, T. L. A computer readability formula designed for machine scoring. Journal of Applied Psychology, 1975, 60 (2), 283-284.

Durr, W. K. Computer study of high frequency words in popular trade juveniles. The Reading Teacher, 1973, 27 (October) 37-42.

Fang, I. E. By computer: Flesch's reading ease score and a syllable counter. Behavior Science, 1968, 13 (May), 249-251.

Felsenthal, H. Readability: Computer utilization. Paper presented at the annual conference of the International Reading Association, Denver, May 1973.

Fox, S. E. Syntactic maturity and vocabulary diversity in the oral language of kindergarten and primary school children. Elementary English, 1972, 49 (April), 489-496.

Golub, L. A., & Kidder, C. Syntactic density and the computer. Elementary English, 1974, 51, 1128-1131.

Harris, A. J. & Jacobson, M. D. Basic Elementary Reading Vocabularies, New York: Macmillan, 1972.

Harris, A. J., & Jacobson, M. D. Predicting twelfth graders' comprehension scores. Journal of Reading 1976, 20 (1), 43-46.

Helm, C. E., & McIver, C. Automated sentence analysis for language-instruction. Computers and the Humanities, 1974, 8, 239-245.

Hopkins, C. J. An Investigation of the Relationship of Selected Oral Language Measures and First-Grade Reading Achievement, Unpublished doctoral dissertation, Purdue University, 1976.

Hopkins, C. J. The spontaneous oral vocabulary of children in grade 1. The Elementary School Journal, 1979, 79 (4), 240-249.

Hopkins, C. J., & Moe, A. J. The validation of a synthetic syllable count appropriate for computer determined readability estimates. Paper presented at the annual conference of the International Reading Association, New York, May 1975. (ED 112 375).

Jacobson, M. D., & MacDougall, M. A. Computer management of information and structure in computer supported instructional materials. Educational Technology, 1970, 10, 39-42.

Klare, G. R., Rowe, P. P., St. John, M. G., & Stolurow, L. M. Automation of the Flesch Reading Ease readability formula with various options. Reading Research Quarterly, 1969, 4 (4), 550-559.

Kidder, C. L., & Golub, L. S. Computer application of a syntactic density measure. Computers and the Humanities, 1976, 10, 325-331.

Kucera, H., & Francis, W. N. Computational Analysis of Present-Day American English. Providence: Brown University Press, 1967.

Mason, G. E., & Blanchard, J. S. Computer Applications in Reading. Newark, Delaware: International Reading Association, 1979.

McCormick, S. H. Syntactic Complexity and Vocabulary Diversity in Two Group's of Children's Books. Unpublished doctoral dissertation, Ohio State University, 1976.

Miller, J. W., & Hintzman, C. A. Syntactic complexity of Newbery award winning books. The Reading Teacher, 1975, 28 (May), 750-756.

Moe, A. J. Word lists for beginning readers. Reading Improvement, 1973, 10, (Fall), 11-15.

Moe, A. J. A comparative study of vocabulary diversity: The speaking vocabularies of first-grade children, the vocabularies of selected first-grade primers, and the vocabularies of selected first-grade trade books. A paper presented at the annual meeting of the American Educational Research Association, Chicago, April 17, 1974. (ED 090 520).

Moe, A. J., & Arnold, R. D. Computer-assisted readability levels of twenty-five Newbery books. Reading Improvement, 1975, 12, 59-64.

Moe, A. J., & Hopkins, C. J. Parsing word strings from text with a computer: Implications for reading instruction. A paper presented at the annual meeting of the International Reading Association, Houston, May 4, 1978.

Moe, A. J., Rush, R. T., & Storlie, R. L. The Literacy Requirements of a Secretary on the Job and in a Vocational Training Program, Purdue University, November 1979.

Patton, M. J., Fuhriman, A. J., & Bieber, M. R. A model and meta-language for research on psychological counseling. Journal of Counseling Psychology, 1977, 24, 25-34.

Pepinsky, H. B. A computer-assisted language analysis system (CALAS) and its applications. ED 162 663, 1978.

Ratkowsky, D. A., & Hantrais, L. Tables for comparing the richness and structure of vocabulary in texts of different lengths. Computers and the Humanities, 1975, 9, 69-75.

Rush, J. E., Pepinsky, H. B., Meara, N. M., Landry, B. C., Strong, S. M., Valley, J. A., & Young, C. E. A computer-assisted language analysis system, Columbus, Ohio: Computer and Information Science Research Center, The Ohio State University, 1974.

Rush, R. T. A program for estimating the readability of English language texts. Unpublished paper, Purdue University, 1979.

Selden, R. W. Use of linguistic string analysis to explore frequency phenomena in the recognition of syntax. Paper presented at the annual meeting of the National Reading Conference, St. Petersburg, Florida, December 1, 1978.

Shamo, G. W. Predicting syllable count by computer. Journalism Quarterly, 1975, 52, 344-346.

Strong, S. M. An algorithm for generating structural surrogates of English text. Journal of the American Society for Information Science, 1974, 25, (1), 10-24.

Thomas, G., Hartley, R. D., & Kincaid, J. P. Test-retest and inter-analyst reliability of the Automated Readability Index, Flesch Reading Ease score, and the Fog count. Journal of Reading Behavior, 1975, 7 (2), 149-154.

Walker, N., & Boillot, M. A computerized reading level analysis. Educational Technology, 1979, 19 (January), 47-49.

Watkins, C. E. Readability, (A program developed by Project CONDUIT). Oregon State University, Corvallis, Oregon. 1975 (approximate date).

Woods, M. L. & Moe, A. J. Analytical Reading Inventory, Columbus: Charles E. Merrill Publishing Company, 1977.